Journal of Data Science **2**(2004), 273-285

An Analysis of Quasi-complete Binary Data with Logistic Models: Applications to Alcohol Abuse Data

Mandy C. Webb[1], Jeffrey R. Wilson[2], and Jenny Chong[3]
[1]*Bureau of Labor Statistics,* [2]*Arizona State University*
and [3]*University of Arizona*

Abstract: This paper examines the issues surrounding the analysis of quasi-complete binary data using logistic regression models with the aid of some popular statistical software programs. Results from three procedures in SAS (LOGISTIC, CATMOD and GENMOD) and the pull-down menu in SPSS were examined. The review was conducted in response to an observation that some users of these procedures do not always independently account for data irregularities encountered when interpreting the computer results. This may be due partly to the fact that the information provided by some statistical software packages may not be sufficient for the user to make informed decisions regarding the results. The dataset that motivated this review came from a substance abuse treatment outcome study. Thirty subjects were followed up to determine the proportion that relapsed and to determine the factors that may predict the relapse. Binary logistic regression models were used to determine the predictors of a relapse. Results showed that there was quasi-complete separation of the data and as such the interpretation is limited. SAS and its procedures in the analysis of quasi-complete data gave very large standard errors, computed more iterations, and provided a useful warning for researchers regarding the configuration of data. In contrast, SPSS provided estimates with smaller standard errors, and did not necessarily provide warning for researchers of the data configuration. Thus researchers who make use of statistical softwares without the knowledge of the iterative procedures used by the statistical package should be aware of the possibility of erroneous conclusions as a consequence when analyzing quasi-complete or complete data.

Key words: Infinite parameters, separation.

1. Introduction

Logistic regression models are commonly used for modeling binary data in clinical, public health, environmental health and epidemiologic studies. For example, logistic regression models have been used to analyze alcohol and drug abuse data for profiling groups of individuals with substance abuse problems.

A logistic regression model can be fitted using SAS and SPSS, among other softwares. Both use iterative procedures involving approximations to obtain parameter estimates for testing hypotheses and predictions. These approximations, though based on asymptotic maximum likelihood methods, can result in incorrect results, depending on the data configuration. Exact methods are available to remedy these problems using other software such as StatXact5 or LogXact 4.1, Oster 2002. However, as these packages are not as readily familiar to researchers when analyzing quasi-complete data with logistic regression models, the guidelines for using two of the more common statistical software packages (SAS and SPSS) should be clearly stated.

The existence, finiteness, and uniqueness of maximum likelihood estimates for the logistic regression model depend on the pattern of data points in the observation space (Albert and Anderson 1984; Santner and Duffy 1986). If the data are completely or partially separated, it may not be possible to obtain reliable maximum likelihood estimates since convergence may not occur. Convergence does not occur because one or more parameters in the model become theoretically infinite. Such is the case if the model perfectly predicts the response or if there are more parameters in the model than can be estimated because the data are sparse. Parameter estimates are usually obtained through solving the normal equations. The maximum likelihood estimates exist only if the normal equations produce a finite solution. Often, a non-unique maximum can occur on the boundary of the parameter space, at infinity.

Several procedures are available for fitting logistic regression models. In SAS, these procedures include the logistic procedure (PROC LOGISTIC), the categorical data modeling procedure (PROC CATMOD), and the generalized linear models procedure (PROC GENMOD). In SPSS there is the "Statistics" pull down menu with options for logistic regression. Criteria for convergence in each of these programs differ slightly. When obtaining the parameter estimates, for example, the LOGISTIC procedure in SAS converges when the largest change among the parameters is small. Hence, convergence does not occur when the parameter is infinite. Eliminating variables can solve the problem of infinite parameters in the case of data separation. However, it is impossible to determine those variables suitable for elimination due to the simultaneous effects.

The existence of a maximum likelihood estimate depends on the concavity of the log likelihood function. However, concavity of the log likelihood function alone does not imply that the maximum likelihood always exists. Silvapulle (1986) obtained a necessary and sufficient condition for the existence of the maximum likelihood estimator for a class of linear regression models for grouped and ungrouped data. This condition has an intuitively simple interpretation. For a given set of data, these conditions may be verified by linear programming.

The problems of existence, uniqueness and location of maximum likelihood estimates in log linear models have received special attention, (Haberman, 1974, Wedderburn, 1976; Silvapulle, 1981). For logistic regression models, the existence theorems fall into three mutually exclusive and exhaustive categories: complete separation, quasi-complete separation and overlap (Albert and Anderson, 1984).

This paper examined the quasi-complete separation in the *Relapse Data* when fitting logistic regression models with the aid of SAS and SPSS. Section 2 reviews quasi-complete separation, complete separation and their properties compared to overlapped data. Section 3 presents the logistic regression model as encountered while modeling the *Relapse Data* with the aid of SAS and SPSS. Some remarks for researchers who may encounter this phenomenon are given in Section 4.

2. Data Configurations for Binary Responses

2.1 Relapse

Data obtained from a substance abuse treatment outcome study were analyzed to determine the influence of selected variables on relapse among individuals who had successfully completed a residential program. For this study, a person who reported using drugs or alcohol within thirty days of discharge is defined as having had a relapse. The predictors considered for selection were age, previous alcohol violations (leading to arrests) and the current severity of family relationship problems. These data (*Relapse Data*) form the reference for our review of quasi- complete data with SAS and SPSS.

The *Relapse Data*, as is true for all other data sets, have to be one of three mutually exclusive and exhaustive categories: complete separation, quasi-complete separation, and overlap. The configuration of the *Relapse Data* shows partial separation, that is, is quasi-complete because one of the variables has a constant value, that is, there was no case that had an alternative value. Of the 30 substance abuse treatment clients followed up after they had discharged from treatment, 16 did not relapse and did not have previous violations related to alcohol or drugs. All patients who had alcohol violations had relapsed; that is, there was no individual with alcohol violations who did not relapse.

2.2 Derived data

The following derived data (Table 1) consist of ten observations with two continuous independent variables with integer values and binary response variable. Let Y_{ij} be a binary response variable taking on the values 1 and 0, where Y_{ij} represents the j th response of the i th subject. Let x_i be the vector of independent explanatory variables, which includes the constant 1, associated with the intercept. Let X_{1i} and X_{2i} for the i th subject be the two corresponding independent variables. Data configurations can be grouped into three mutually exclusive and exhaustive categories: complete separation, quasi-complete separation, and overlap.

Complete separation

Complete separation of data points exists if there is a vector b that correctly allocates all observations to their appropriate response group; that is, $b'x_i > 0$ for $Y_{ij} = 1$ and $b'x_i < 0$ for $Y_{ij} = 0$. When a data set is completely separated, there are non-unique infinite estimates. Allowing the iterations of the maximizing likelihood function to continue, one can see that the log likelihood approaches zero. The dispersion matrix becomes unbounded, (So 1993). A plot of the data in Table 1 is shown below in Figure 1.

Table 1: Derived data for configuration

Observation	Y_{ij}	$\mathbf{X}_{1i}$	$\mathbf{X}_{2i}$	Observation	Y_{ij}	$\mathbf{X}_{1i}$	$\mathbf{X}_{2i}$
De-nosied	Absolute	Percent					
1	0	15	40	6	1	35	40
2	0	30	45	7	1	27	28
3	0	34	50	8	1	15	20
4	0	18	49	9	1	38	30
5	0	27	41	10	1	29	20

The complete separation is apparent when one plots the data, that is, there is no overlap between the $Y_{ij} = 1$ and $Y_{ij} = 0$ observations.

Quasi-complete separation

Quasi-complete separation occurs when there exists a vector b such that $b'x_i \geq 0$ for all j such that $Y_{ij} = 1$ and $b'x_i \leq 0$ for all j such that $Y_j = 0$ (***** Do you mean Y_{ij}? *****). The equality must hold for at least one subject in each group. As with complete separation, this configuration also yields non-unique infinite estimates. If the iterative process of maximizing the likelihood function is allowed to continue, the dispersion matrix becomes unbounded as in completely separated data sets. However, as the log likelihood diminishes to zero in the completely separated case; it approaches a nonzero constant when a data set is quasi-completely separated. Consider the data in Table 1 with one small change. The independent values of the first observation are $X_{1i} = 30$ and $X_{2i} = 35$.

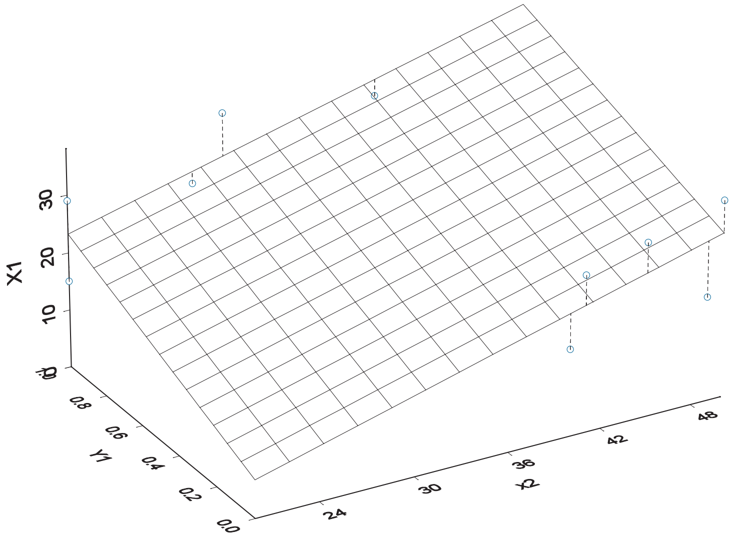

Figure 1: Completed separated

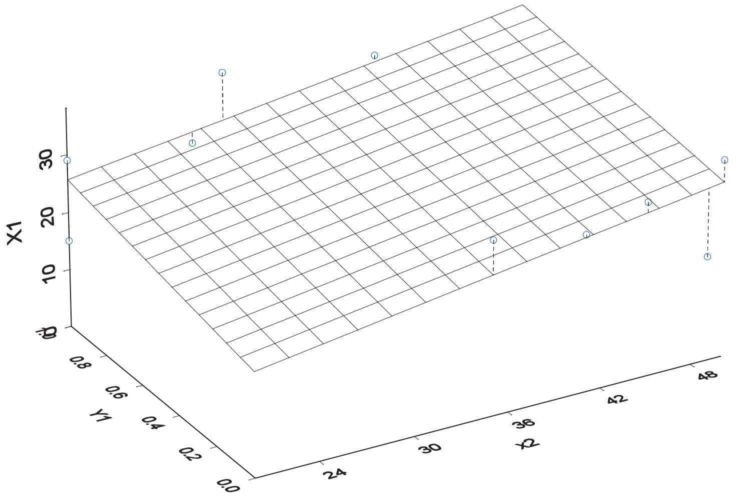

Figure 2: Quasi-completed separated

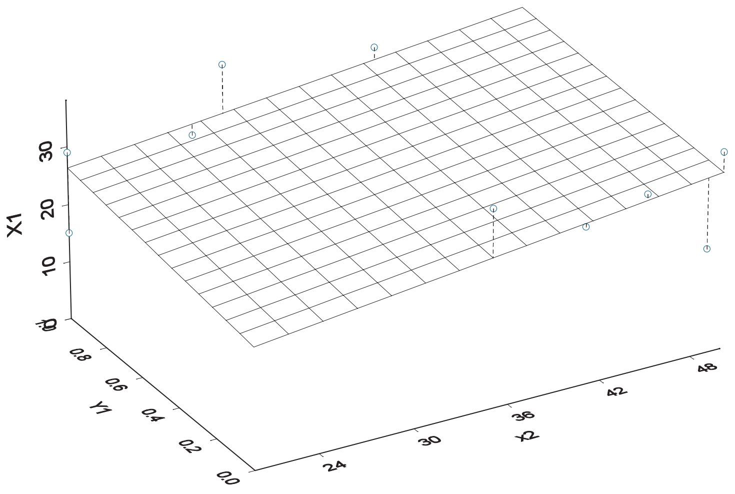

Figure 3: Overlapped

Figure 2 shows a plot of the data with the first observation replaced. An examination of the graph of the data points show that we can define a non-trivial vector such that the vector separates the data points into their respective response group with at least one of each response group lying directly on the line defined by the vector of the form (X_1, X_2, β) is $(1, -1, 5)$. The line $X_2 = X_1 - 5$ separates the data in two dimensions, but has one observation from each response lying on the line.

A typical characteristic of quasi-complete separation is that the variances of the pseudo estimates are seemingly large. Generally, quasi-complete separation arises only when a number of the explanatory variables have integer coefficients, and these integral values are forced by the structure of the problem.

Overlapped

When there is no separation (complete or quasi-complete) found in the data configuration, there is said to be an overlap of sample points. In this data configuration, the maximum likelihood estimates exist and are unique. Generally, complete separation and quasi-complete separation are problems typically associated with small data sets. Complete separation can occur with any type of data; however quasi-complete separation rarely is present with truly continuous explanatory variables. Recall the data in Table 1 and replace the first observation with the values of $X_{1i} = 34$ and $X_{2i} = 35$. The data are now in an overlapped configuration. This is also seen in Figure 3.

2.3 Binary data configurations with statistical software

The *Relapse data* as described in this paper were found to be quasi-complete. It was the motivating factor for looking at data configurations. When looking at a graph of the data, we found a non-trivial vector such that the vector separates the data points into their respective response group with two; one of each response group lying directly on the line defined by the vector.

If the data configuration is one of complete separation, it will be apparent graphically since no overlap will be seen between the cases when $Y_{ij} = 1$ and $Y_{ij} = 0$. Using SAS procedures when fitting logistic regression models would have resulted in a warning: "*There is a complete separation of data points and that the maximum likelihood estimate does not exist*". In particular, the warning says: "*The LOGISTIC procedure continues in spite of this warning. The results given are based on the last maximum likelihood iteration. Validity of the model fit is questionable.*" Similar warnings are obtained using PROC GENMOD in SAS. The warning states: "*Convergence is not attained for at least one side of profile likelihood confidence interval, the number of iterations i equals 50.*" Thus the researcher of the *Relapse data* would be alerted that the data are separated and do not provide conclusive evidence regarding the impact of the predictors. This configuration also yielded non-unique infinite estimates. If the iterative process of maximizing the likelihood function were allowed to continue, the dispersion matrix becomes unbounded.

When confronted with quasi-complete data as it is in the *Relapse Data*, PROC LOGISTIC gave the warning: "*That there is possibly a quasi-complete separation of data points. In particular, the maximum likelihood estimate may not exist. The LOGISTIC procedure continues in spite of the above warning. Results shown are based on the last maximum likelihood iteration. Validity of the model fit is questionable.*" It is typical of quasi-complete separation data that the variances of the pseudo estimates are seemingly large. A slight perturbation of appropriate data points would remove the separation, or make it complete. Such a situation can be tested by appropriately perturbing the coefficients by a percent or so (not any coefficients which have known exact values). If the answers vary wildly, this suggests numerical ill conditioning, and the specific solution has limited meaning.

If the *Relapse Data* were complete there would have been a vector b that correctly allocates all observations to their appropriate response group; that is, $b'x_i > 0$ for $Y_{ij} = 1$ and $b'x_i < 0$ for $Y_{ij} = 0$ where Y_{ij} represents the j th response of the i th subject and x_i be the vector consisting of age, alcohol violations, family relationship, and the constant 1 associated with the intercept. A complete data set it would have produced non-unique infinite estimates, with the log likelihood approaching zero and the dispersion matrix becoming unbounded, (So, 1993). Webb (2002) reported that there are some shortcomings of the definition for complete separation since for $b = 0$ the restriction will always hold. Therefore, one needs a nonzero vector, which none of the software tutorials explicitly mention. In particular, one needs to determine if there is a nonzero feasible vector. Silvapulle and Burridge (1986) demonstrated one such method for determination.

If the *Relapse Data* were overlapped, then there would have been cases in the two response groups distributed across the categories of the response variable and the maximum likelihood estimates would exist and be unique. Complete separation and quasi-complete separation are problems generally associated with small data sets. Complete separation can occur with any type of data; however quasi-complete separation rarely is present with truly continuous explanatory variables.

3. Binary Logistic Models

There are several statistical procedures that are useful for fitting binary logistic regression models. In particular in SAS, the procedures include PROC LOGISTIC, PROC CATMOD, and PROC GENMOD and in SPSS there is a pull-down menu for binary logistic regression.

A logistic regression model was fitted to the *Relapse Data* with the binary response as whether or not a person used drugs and/or alcohol within the 30 days of discharge. Age, alcohol violation (ALCVIOL), and a family relationship score (FAMCOMP) were used as prognostic factors. Alcohol violation is a binary variable, where 0 denoted no arrests and 1 denoted at least one arrest. The family score ranged in values from 0.00 to 1.00, where 0.00 means that there were no family problems and 1.00 indicating severe family problems. The variable age was measured in years.

When fitting logistic regression models with SAS using PROC LOGISTIC on the *Relapse Data*, the estimation process does not converge since the largest change among the parameters is small. In this case, convergence cannot be obtained since one or more of the parameters are infinite. In such situations one may want to reduce the number of variables and/or recode the continuous variables so they are treated as categorical. However, as many researchers have pointed out, there is no easy way to know exactly which variables should be eliminated or which continuous variables to categorize as the variables are fitted simultaneously. Instead of reducing or altering the variables, one can use a different rule to decide when to stop the iterations. Such adjustment is possible when using SAS as shown i n SAS 6.11 release LOGISTIC documentation. The resulting model is usually appropriate for prediction or classifying observations, but not suitable

Table 2: Iteration history for PROC LOGISTIC for relapse data

Iteration	$-2\log L$	Intercept	AGE	FAMCOMP	ALCVIOL
0	41.46	−0.1335	0	0	0
1	21.59	−0.453454	−0.038254	3.363752	2.574144
2	17.33	0.918579	−0.108884	6.079823	4.107869
3	14.66	3.138064	−0.217551	9.730740	6.351777
4	12.93	6.081887	−0.372817	15.504767	9.658252
5	12.25	8.930162	−0.528479	21.797471	13.132423
6	12.12	10.546430	−0.617270	25.561036	15.535184
7	12.10	10.925740	−0.637943	26.434712	16.861932
8	12.10	10.941250	−0.638783	26.469832	17.877148
9	12.09	10.941274	−0.638784	26.469888	18.877988
10	12.09	10.941274	−0.638784	26.469888	19.878290
11	12.09	10.941274	−0.638784	26.469888	20.878401
12	12.09	10.941274	−0.638784	26.469888	21.878442

to make inferences.

The logistic regression procedure in SAS has a fairly simple approach to recognize data configurations that lead to infinite parameter estimates. The idea behind this empirical approach is that any convergence method of maximizing the log likelihood will necessarily yield a solution giving complete separation when such a solution is possible. Within the first eight iterations, if SAS had met convergence criteria in maximizing the log likelihood function, SAS does not perform an internal check for complete or quasi-complete separation. If the convergence criterion is not met within the first eight iterations, as with the *Relapse data*, SAS computes the probabilities of each observation's observed response. Complete separation occurs when the probability of all observations' observed response is one or zero. When this occurs the iteration process ends. If the data configuration is not completely separated, the SAS program then looks for an observation having a large probability (> 0.95) of the observed response. Since the *Relapse data* have a quasi-complete separated configuration, the asymptotic dispersion matrix was not bounded and some of the diagonal elements of the dispersion matrix were greater than or equal to 5000. Under these conditions, the iteration process subsequently ended and the SAS program stated that the data configuration is quasi-complete.

When PROC LOGISTIC detects a quasi-complete separation, which impacts on the validity of the model, SAS produces a warning that the maximum likelihood estimates may not exist because quasi-complete separation of data points is detected. The iteration history for the *Relapse data* is given in Table 2. SAS

Table 3: Parameter estimates, standard errors, and P-values

Parameter	SAS (LOGISTIC)	SAS (CATMOD)	SPSS
Intercept	10.941	−21.373	10.941
	8.015	8.001	8.014
	0.172	0.008	0.172
AGE	−0.639	0.639	−0.639
	0.407	0.406	0.407
	0.116	0.116	0.116
FAMCOMP	26.470	−26.462	26.470
	16.593	16.563	16.593
	0.111	0.110	0.111
ALCVIOL	21.878	0.000	19.872
	182.50	10.436	67.173
	0.905	......	0.767

continues processing the data and gives the estimates and standard errors, Table 3.

The LOGISTIC procedure in SAS offers two iterative maximum likelihood algorithms: The Fisher-scoring method (default algorithm) and the Newton-Raphson method. Fisher-scoring method is the same as fitting a model by iteratively reweighting least squares. Both algorithms, Fisher-scoring and Newton-Raphson, will give the same parameter estimates in this instance. However, the estimated covariance matrices of the parameter estimators are not necessarily the same. This is due to the fact that the Fisher-scoring method is based on the expected information matrix while the Newton-Raphson method is based on the observed information matrix (SAS Inc., *Iterative Algorithms for Model-Fitting*). In the fit of a binary logistic model the observed and expected information matrices are the same resulting in identical estimated covariance matrices.

The *Relapse data* were also fitted using the GENMOD procedure, which fits a generalized linear model to the data by maximum likelihood estimation of the vector of parameters. In general, there is no closed form solution for the maximum likelihood estimates of the parameters. The PROC GENMOD, as with the other SAS procedures, uses an iterative process to estimate the parameters. When infinite parameters are present in the model, it signifies that either there are one or more zero frequencies or there is a poor model choice with collinearity among the estimates.

In SPSS, the maximum likelihood estimates for the vector of parameter are obtained using a Newton-Raphson based algorithm. The convergence of the maximum likelihood estimates can be based on the following: the absolute difference for the parameter estimates between iterations, the percent difference in the log-likelihood function between successive iterations, or the maximum number of iterations specified. Otherwise, if during the iterations, the value of the product of the predicted probabilities and its complement is less than 10^{-10} for all cases then the log-likelihood function is very close to zero. When this occurs, the iterations stops and the message "All predicted values are either 1 or 0" is displayed.

In light of these approaches, when using SPSS neither age, family score, nor alcohol violations, had a significant effect on the probability of a person discharged from rehabilitation center relapsing within thirty days. Although none of the variables were found to be significant in the model, it is of interest to examine the data configuration and check whether the maximum likelihood estimates converge. The initial -2 log likelihood has a value of 41.455. The maximum likelihood estimates converged on the tenth iteration. The last decrease in the log-likelihood function was less than .001

The approach to the analysis of quasi-complete data may differ among programs in the iterative procedures. For example, SAS terminates at the twelfth iteration whereas a program such as SPSS may terminate at the tenth iteration. An examination of the iteration shows that the last change in the ALCVIOL variable was one unit, while the other variables had negligible change. The last change in the log likelihood function in SAS is 0.0001033108, which does not reach the minimum level for SAS to see convergence in the maximum likelihood functions, which is 10^{-8}. None of the three variables under consideration were found to be significant in predicting the probability that a person will use drugs or alcohol within 30 days out of rehabilitation. SAS detected quasi-complete separation, and the standard error for the ALCVIOL variable is very large. The estimates for the variable ALCVIOL differ substantially in both programs.

The data were also fitted using PROC CATMOD and PROC GENMOD in SAS. The CATMOD procedure stated that the maximum likelihood computations converged but noted that ALCVIOL may be a redundant or restricted parameter. It refers to the parameter estimate associated with ALCVIOL as infinite. The estimates based on PROC CATMOD are given in Table 3.

The GENMOD procedure also produced a warning about the convergence of the log maximum likelihood function: "*that the negative of the Hessian is not positive definite and the convergence is questionable. The procedure continues but the validity of the model fit is questionable*". The specified model did not converge.

4. Conclusions

When fitting binary logistic regression models using SPSS, or SAS with LO-

GISTIC, CATMOD or GENMOD procedures, the resulting parameter estimates may differ substantially if the data are not overlapped . In particular SAS acknowledges or notifies the user of the potential validity of the model, while SPSS may not provide such clarity in a warning. This is in part due to the fact that SPSS stops at the 10th iteration while SAS goes to the 12th and beyond if the "NO CHECK" option is instituted in the program. In these cases when quasi-complete or complete data are encountered the standard errors provided by SPSS may be substantially smaller than the standard errors provided by SAS. However, these standard errors are not appropriate to be used in any analysis.

The *Relapse data* have a quasi-complete data configuration. Since SAS delivered warnings regarding the configuration of the data, the researcher should use this as an alert to look for alternate models. Not all programs will necessarily give these details and as such the researcher may conclude that the variables are all non- significant when in fact that is not known. In particular the results give the impression that alcohol violations variable was not significant when it is impossible to determine this due to the quasi-complete separation between relapse and alcohol violations. These results suggest that not all softwares are equipped for binary logistic models when the dataset is sparse and there are sampling zeros and more importantly the data is not overlapped. Indeed, since it is possible for software programs to fail to acknowledge the quasi-complete separation, the use of any statistical packages without caution under irregular conditions is ill advised. But then how does one know that the conditions are irregular.

Even if there is no alert from the statistical software regarding data separation one should be concerned if the standard errors are large. Omitting a variable and re-running the data is one method of approach. If the coefficients change substantially then this is a further sign to explore the data configurations.

References

Albert, A. and Anderson, A. (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika* **71**, 1-10.

Haberman, S. J. (1974). *The Analysis of Frequency Data*. University of Chicago Press.

Oster, R. (2002). An examination of statistical software packages for categorical data analysis using exact methods. *The American Statisticians* **56**, 235-356.

Santner, T. J., and Duffy, D. E. (1986). A note on A. Albert and J. A. Anderson's conditions for the existence of maximum likelihood estimates in logistic regression models. *Biometrika* **73**, 755-758.

Silvapulle, M. J., (1981). On the existence of maximum likelihood estimators for the binomial response models. *Journal of the Royal Statistical Society, Series B* **43**, 310-313.

Silvapulle, M. J., and Burridge, J. (1986). Existence of maximum likelihood estimates in regression models for grouped and ungrouped data. *Journal of the Royal Statistical Society, Series B* **48**, 100-106.

So, Y., (1993). A tutorial on logistic regression. *Proceedings of the SAS Users Group International Conference* **18**, 1290-1295.

SAS Institute Inc. *SAS OnlineDocR, Version 8.* The CATMOD, LOGISTIC, GENMOD procedures (1999). SAS Institute.

SAS Institute Inc. *SAS OnlineDocR, Version 8.* FAQ #960 (1999). SAS Institute.

Webb, M. (2002). Quasi-complete with binary data – MS Paper – Arizona State University.

Wedderburn, R. W. M. (1976). On the existence and uniqueness of the maximum likelihood estimates for certain generalized linear models. *Biometrika* **63**, 27-32.

Received Feburary 25, 2003; accepted August 3, 2003.

Mandy C. Webb
Bureau of Labor Statistics
Washington D.C. 20212
USA

Jeffrey R. Wilson
School of Health Administration and Policy
W.P. Carey School of Business
Arizona State University
Tempe, Arizona 85287-4206
USA
jeffrey.wilson@asu.edu

Jenny Chong
Division of Epidemiology and Biostatistics
Mel and Enid Zuckerman College of Public Health
University of Arizona
Tucson, Arizona 85721-210228
USA